2001

# Using Automated Medical Records for Rapid Identification of Illness Syndromes (Syndromic Surveillance): The Example of Lower Respiratory Infection

Kenneth Kleinman
*University of Massachusetts Amherst*, kkleinman@schoolph.umass.edu

Ross Lazarus

Inna Dashevsky

Alfred DeMaria

Richard Platt

BMC Public Health

Research article

# Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection

Ross Lazarus [1,2], Ken P Kleinman [3], Inna Dashevsky[3], Alfred DeMaria [4] and Richard Platt MD*[1,3]

Address: [1]Chanming Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [2]Department of Public Health and Community Medicine, University of Sydney, Sydney, Australia, [3]Department of Ambulatory Care and Prevention, Harvard Medical School, Harvard Pilgrim Health Care, and Harvard Vanguard Medical Associates, Boston, MA, USA and [4]Bureau of Communicable Disease Control, Massachusetts Department of Public Health, Boston, MA, USA

E-mail: Ross Lazarus - ross.lazarus@channing.harvard.edu; Ken P Kleinman - ken_kleinman@harvardpilgrim.org; Inna Dashevsky - inna_dashevsky@harvardpilgrim.org; Alfred DeMaria - alfred.demaria@state.ma.us; Richard Platt* - richard.platt@channing.harvard.edu

*Corresponding author

Published: 22 October 2001

*BMC Public Health* 2001, **1**:9

Received: 3 August 2001
Accepted: 22 October 2001

This article is available from: http://www.biomedcentral.com/1471-2458/1/9



## Abstract

**Background:** Gaps in disease surveillance capacity, particularly for emerging infections and bioterrorist attack, highlight a need for efficient, real time identification of diseases.

**Methods:** We studied automated records from 1996 through 1999 of approximately 250,000 health plan members in greater Boston.

**Results:** We identified 152,435 lower respiratory infection illness visits, comprising 106,670 episodes during 1,143,208 person-years. Three diagnoses, cough (ICD9CM 786.2), pneumonia not otherwise specified (ICD9CM 486) and acute bronchitis (ICD9CM 466.0) accounted for 91% of these visits, with expected age and sex distributions. Variation of weekly occurrences corresponded closely to national pneumonia and influenza mortality data. There was substantial variation in geographic location of the cases.

**Conclusion:** This information complements existing surveillance programs by assessing the large majority of episodes of illness for which no etiologic agents are identified. Additional advantages include: a) sensitivity, uniformity and efficiency, since detection of events does not depend on clinicians' to actively report diagnoses, b) timeliness, the data are available within a day of the clinical event; and c) ease of integration into automated surveillance systems.

These features facilitate early detection of conditions of public health importance, including regularly occurring events like seasonal respiratory illness, as well as unusual occurrences, such as a bioterrorist attack that first manifests as respiratory symptoms. These methods should also be applicable to other infectious and non-infectious conditions. Knowledge of disease patterns in real time may also help clinicians to manage patients, and assist health plan administrators in allocating resources efficiently.

## Introduction

Public health agencies, medical care delivery systems, and clinicians all depend on accurate and timely information about disease occurrence to guide planning, resource allocation, and case management. In the public health arena, the current United States infectious disease surveillance infrastructure was recently the subject of an extensive review by the US Department of Health and Human Services [1]. That review identified substantial weaknesses in existing capacity to detect four major threats to public health – emerging infections, antimicrobial resistance, bioterrorism, and pandemic influenza. A major recommendation of that report was to engage with the health care delivery system as a partner in surveillance. Specifically, it made a "level 1" recommendation for funding of efficient, easy-to-use, and rapid automated reporting systems based on national standards. Such automated reporting systems can take two basic forms – they can facilitate clinicians' active reporting of individual cases of diseases of interest, or they can use some of the extensive information that is already collected in automated form in the process of medical care delivery or in administration of medical care benefits.

We focus here on an example of the latter approach, assessing automated information about diagnoses assigned by clinicians during ambulatory care visits. To our knowledge, automated medical record information has not been used for rapid surveillance purposes. Such an approach has several potential advantages. These include the facts that it imposes no reporting burden on clinicians and avoids the reporting biases inherent in spontaneous reports, and the data can be available promptly. In addition, they can assess the very large number of cases of illness for which no etiologic diagnosis exists, and which are therefore typically not reportable. These illness syndromes can be of interest because they might provide several days' advance warning of a serious problem, such as anthrax, and also because they make use of the large body of information in encounters for which no diagnostic testing is performed. Currently, this information is typically unavailable for epidemiologic purposes, although the tracking of influenza like illness in sentinel practices is an example of such syndromic surveillance.

The automated medical records of health plans and large group practices are an especially useful source of surveillance data because they serve a well-defined population of members, they have responsibility for most aspects of health care, and they routinely collect clinical, demographic and accounting data. This paper describes some of the technical and methodological issues encountered in developing a surveillance system for lower respiratory infection based on automated ambulatory care electronic encounter records from a large HMO and multi-specialty group practice.

## Methods

We studied information in the automated medical records, demographic, and eligibility records of individuals cared for by Harvard Vanguard Medical Associates (HVMA), a large multi-specialty group practice in eastern Massachusetts. At the beginning of the study period, HVMA was a staff model component of Harvard Pilgrim Health Care (HPHC), a not-for-profit HMO, and all individuals studied were members of Harvard Pilgrim. Ambulatory care, including scheduled, same day, and urgent care visits, is delivered in fourteen health centers in greater Boston. Urgent care visits include many of the encounters that are cared for in hospital emergency rooms in other practice settings. All individuals included in this study were HMO members who had a strong financial incentive to receive their care at one of these health centers.

Care in the health centers is delivered using automated medical records. At present, Epicare, a commercially available system is used. Clinicians assign diagnoses for each visit from lists that correspond to ICD9CM codes [2]. The system also contains vital signs and all providers' full text notes. An earlier automated medical record system in use during the first part of the study period used COSTAR[3] diagnosis codes, which we mapped to ICD9CM codes. Individual patient identifiers were replaced with study-specific, encrypted unique identifiers before the data were made available for analysis. The study was approved by the HPHC Institutional Review Board.

A total of 7,265,523 encounter records, including routine, scheduled, and urgent care visits, plus telephone calls that clinicians chose to record, were collected during the four years between the beginning of 1996 and the end of 1999. Each encounter was coded by the physician at the time of consultation with up to 11 diagnosis codes. The proportion of visits assigned multiple ICD9CM codes decreased steeply after the first two of these, ranging from 10.2% (744,290) of visits with three codes and 7.4% (540,049) of visits with four codes, to 0% (0) with eleven. Since we were interested in a primary diagnosis of lower respiratory infection, we considered only the first two diagnostic codes. Emergency room visits were excluded for two primary reasons: they are not uniformly captured, and they enter the system after some delay, thereby greatly diminishing their utility for any real time surveillance system.

Membership data for the same period were extracted from administrative data systems. The number of days of

**Table 1: Age and sex distribution of lower respiratory infection encounters and person-time.**

| Age group | Episodes | | Years | | Rate | | Rate ratio (M/F) | 95% CI |
|---|---|---|---|---|---|---|---|---|
| | F | M | F | M | F | M | F | M |
| <1 | 3304 | 4252 | 32060.5 | 33032.2 | 0.1031 | 0.1287 | 1.25 | 1.19–1.31 |
| 1–4 | 6572 | 7523 | 25051.3 | 26424.2 | 0.2623 | 0.2847 | 1.09 | 1.05–1.12 |
| 5–14 | 9860 | 10517 | 76412.1 | 79600.7 | 0.1290 | 0.1321 | 1.02 | 1.00–1.05 |
| 15–24 | 4658 | 3495 | 57973.5 | 44746.9 | 0.0804 | 0.0781 | 0.97 | 0.93–1.01 |
| 25–34 | 7029 | 3841 | 135954.1 | 113350.0 | 0.0517 | 0.0339 | 0.66 | 0.63–0.68 |
| 35–44 | 8896 | 5690 | 120094.0 | 110367.8 | 0.0741 | 0.0516 | 0.70 | 0.67–0.72 |
| 45–54 | 7540 | 5162 | 79264.2 | 74378.0 | 0.0951 | 0.0694 | 0.73 | 0.70–0.76 |
| 55–64 | 4544 | 3105 | 34273.8 | 32041.1 | 0.1326 | 0.0969 | 0.73 | 0.70–0.76 |
| 65–74 | 8654 | 7138 | 26123.8 | 20967.0 | 0.1409 | 0.1154 | 0.82 | 0.79–0.85 |
| 75–84 | 2205 | 1387 | 10840.1 | 7312.5 | 0.2034 | 0.1897 | 0.93 | 0.87–1.00 |
| >85 | 596 | 393 | 2023.0 | 917.0 | 0.2946 | 0.4286 | 1.45 | 1.28–1.65 |

membership over the four years of the study for each individual member was determined. These were summed by age and gender and used as the denominators for calculating crude and age/gender specific incidence rates.

ICD9CM codes were grouped into syndromes for the purposes of the main analysis. A provisional set of ICD9CM codes for Lower Respiratory Infection (LRI) currently in use (personal communication, J. Pavlin; Appendix 1) by the US Department of Defense ESSENCE project [4] were used. Any encounter record with one of the relevant codes in either of the first two ICD9CM codes assigned during an encounter was selected for analysis.

After selecting all LRI records, encounters for each individual were grouped into episodes of illness on the assumption that for any individual patient, a subsequent visit for LRI within six weeks of a preceding LRI visit would be likely to represent follow-up for the same infection, whereas widely separated LRI visits probably represented separate events. The data supporting the six week cut off is shown below.

After these analyses were completed, the health plan discovered that some clinical data were missing in master files from which the analysis datasets were created. Membership information was not affected by this problem. Corrected data were not expected to be available for several months. Analysis of a random sample of 500 individuals' data showed that the results reported here undercount individuals with LRI, LRI episodes, and LRI encounters by less than 10%, with no identified pattern to the missing information. The data are therefore presented with the understanding that complete figures would include an additional 10% of counts; this difference should not affect the substantive conclusions.

## Results

There were 501,323 individuals who were members for at least part of the four year study period and thus eligible to present for ambulatory care. The median duration of membership was 2.2 years (mean = 2.3 years, s.d. = 1.5 years), yielding 1,143,208 person-years of observation. The age and sex distribution of observation time is shown in Table 1.

During this time, 152,435 ambulatory care encounters were assigned one or more of the codes in the LRI syndrome definition (see Appendix 1). Although 119 different ICD9CM diagnoses contribute to the lower respiratory syndrome category, three codes accounted for more than 90% of all encounters (Table 2). These codes were cough (ICD9CM 786.2), pneumonia not otherwise specified (ICD9CM 486), and acute bronchitis (ICD9CM 466.0). Notably, these are codes that clinicians can assign without performing laboratory testing.

The 152,435 LRI encounters were attributable to 75,747 individual members who experienced a mean of 2.0 encounters each (minimum = 1, maximum = 65). More than half of these individuals (n = 43,404) had exactly one LRI encounter. For the 32,343 individuals with more than one LRI encounter, the distribution of intervals between LRI encounters is shown in Figure 1. In more than half of individuals with multiple LRI encounters, a sec-

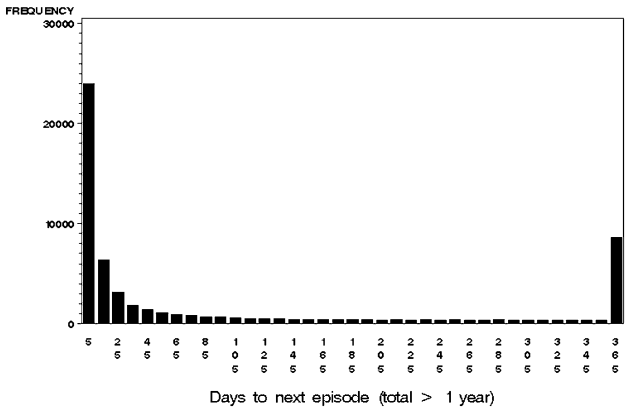


**Figure 1**
Frequency distribution of intervals between multiple LRI encounters for individual members with more than one encounter over the study period.

**Table 2: Most common ICD9CM diagnoses in lower respiratory infection encounters. Individual episodes can have two of these diagnoses.**

| ICD9CM | Description | Count | Percentage of Total |
|---|---|---|---|
| 786.2 | Cough | 62634 | 52.8 |
| 486 | Pneumonia, organism not otherwise specified | 27681 | 23.4 |
| 466.0 | Acute bronchitis | 18286 | 15.4 |
| 466.1 | Acute bronchiolitis | 5594 | 4.7 |
| 487.1 | Influenza with respiratory manifestations, not elsewhere classified | 1902 | 1.6 |

ond encounter occurred less than three weeks after the first. Approximately 8,000 members had a second encounter more than one year after the first. To avoid double counting of ambulatory care encounters that were really part of the same episode of infection, we used a criterion based on clinical experience together with the distribution shown in Figure 1. If any individual member had more than one LRI encounter, we required 6 weeks free of any LRI diagnosis encounter before attributing a new episode of LRI illness to that individual. Using this criterion, the 152,435 encounters for LRI were reduced to 106,670 distinct episodes of lower respiratory infection, giving an overall annual incidence rate of episodes of LRI coming to medical attention of 93/1,000 person-years.

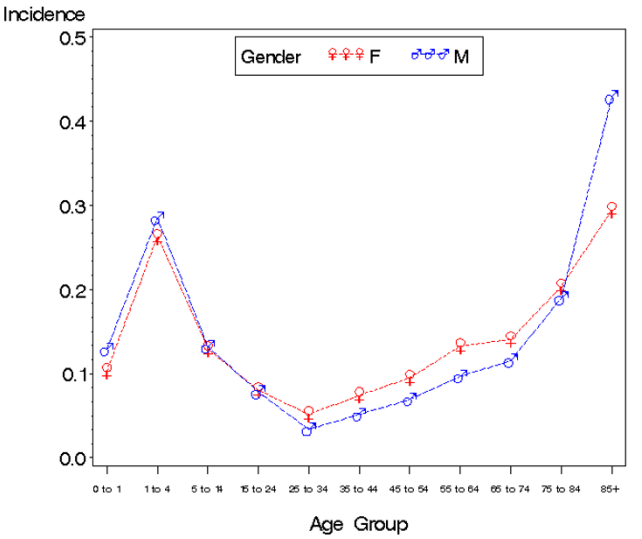


**Figure 2**
Incidence rate (events/person-year) of lower respiratory infection episodes diagnosed in ambulatory care records.

Annual LRI episode incidence rates are shown in Table 1 and Figure 2 by age group and gender. These incidence rates are highest among pre-school children after the first year of life and among those 85 years and older. In those categories, the rates for boys and men exceeded those for girls and women, by 25% and 45%, respectively. In contrast, among adults from 25 through 74 years, the rates among women were 18% to 34% higher than those in men, with the largest excess in the younger age groups. The confidence intervals for these rate ratios excluded the null in each of the age categories, except 5–24 and 75–84 years.

As expected, there were many more LRI episodes during the winter. The weekly counts of these episodes are shown in Figure 3. The seasonal variation in these counts is strikingly similar to the variation in deaths from pneumonia and influenza in 122 cities, reported by the CDC [5], also shown in Figure 3. For most winters, there is a suggestion that the LRI episode counts rise shortly before the peak in deaths.

Because an important surveillance goal is to detect certain events, such as an anthrax exposure, as rapidly as possible, we show daily, rather than weekly, counts in Figure 4. The left hand panel, with over 1,400 data points, also makes evident the overall seasonal variation in disease incidence, though with considerably more scatter. This figure also shows that diagnoses are much less likely to be assigned on weekend days. The right hand panel, which shows daily counts for a single month demonstrates more clearly the day to day variation in oc-

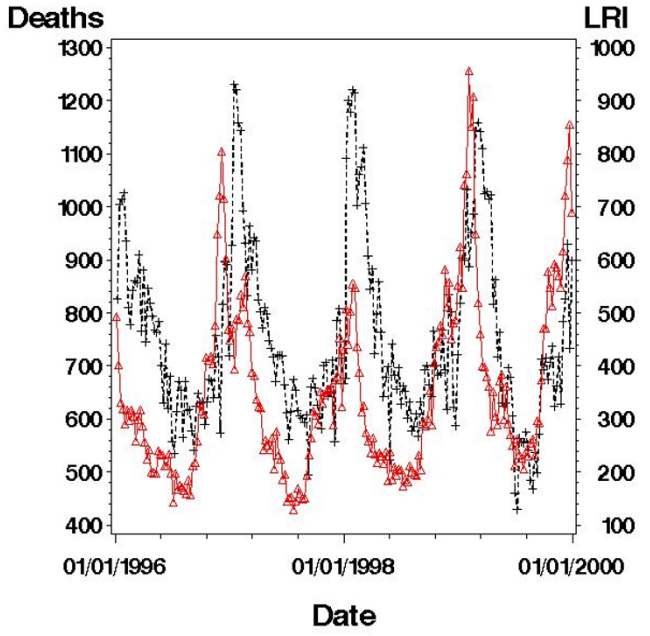


**Figure 3**
Weekly counts of pneumonia and influenza deaths reported to CDC from U.S. cities, including greater Boston (dashed black line, left hand vertical axis) and lower respiratory infection episodes diagnosed in greater Boston ambulatory settings (solid red line, right hand vertical axis).

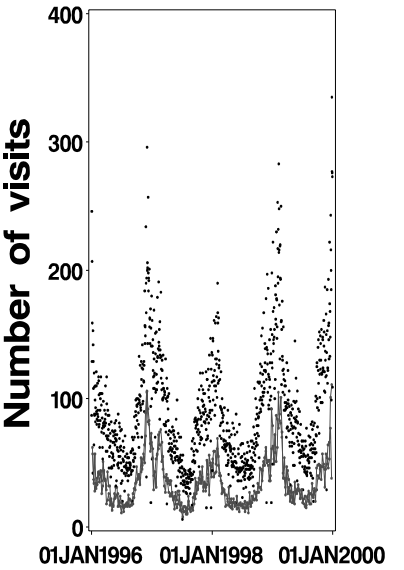


**Figure 4**
Daily counts of lower respiratory infection episodes. The left hand panel shows all years' data. Counts for Monday through Friday are shown in black; counts for Saturdays and Sundays are shown in red. The right hand panel shows the same data for a single month.

currence of LRI episodes. The lower weekend counts are seen, usually followed by the week's highest counts on Mondays. However, the overall day to day variation in counts is relatively large, making it difficult to discern important overall trends in disease incidence. This difficulty is evident even in this best case situation in which all of the events in greater Boston are aggregated.

In Figure 5 we show disease patterns in space as well as time. Episodes are assigned to the census block group in which the individual resides. The four panels in the figure show intervals ranging from the entire four years to a single day. For intervals as small as a week, the overall geographic clustering of episodes, corresponding to the concentration of residents who receive care from Harvard Vanguard, is reasonably well preserved. The small number of events in any locale is well demonstrated, however. A similar display of rates, rather than counts would show much less variation.

## Discussion

Routinely collected medical record information has several important advantages for surveillance purposes, and particularly for the surveillance of infection syndromes. Among the most important is the ability to assess the large majority of episodes of illness for which no etiologic agents are identified, either because good medical practice does not require that clinicians perform diagnostic tests, or because an unusual agent may fail to be detected by standard laboratory tests. Syndromic surveillance should thus be useful both for tracking the activity of infectious agents that are common in communities, and also for identification of new, emerging infections or bioterrorist attack. The detection of an increased frequency of events would typically trigger more intensive assessment, including more diagnostic testing than would ordinarily be indicated. Syndromic surveillance also allows the earliest possible identification of increased disease frequency, presumably days before laboratory test results become available. This early indication of a problem may be important in detecting and responding to a bioterrorist attack, for instance the release of anthrax in a community.

Other advantages of automated diagnosis data for surveillance include uniformity and increased sensitivity of detection, since clinicians are not required to recognize a condition as being of interest. These data also circumvent the need for providers to initiate reporting, an important consideration in light of the time pressures that constrain existing reporting. For some purposes, automated methods may augment or replace resource intensive sentinel surveillance programs, for example those

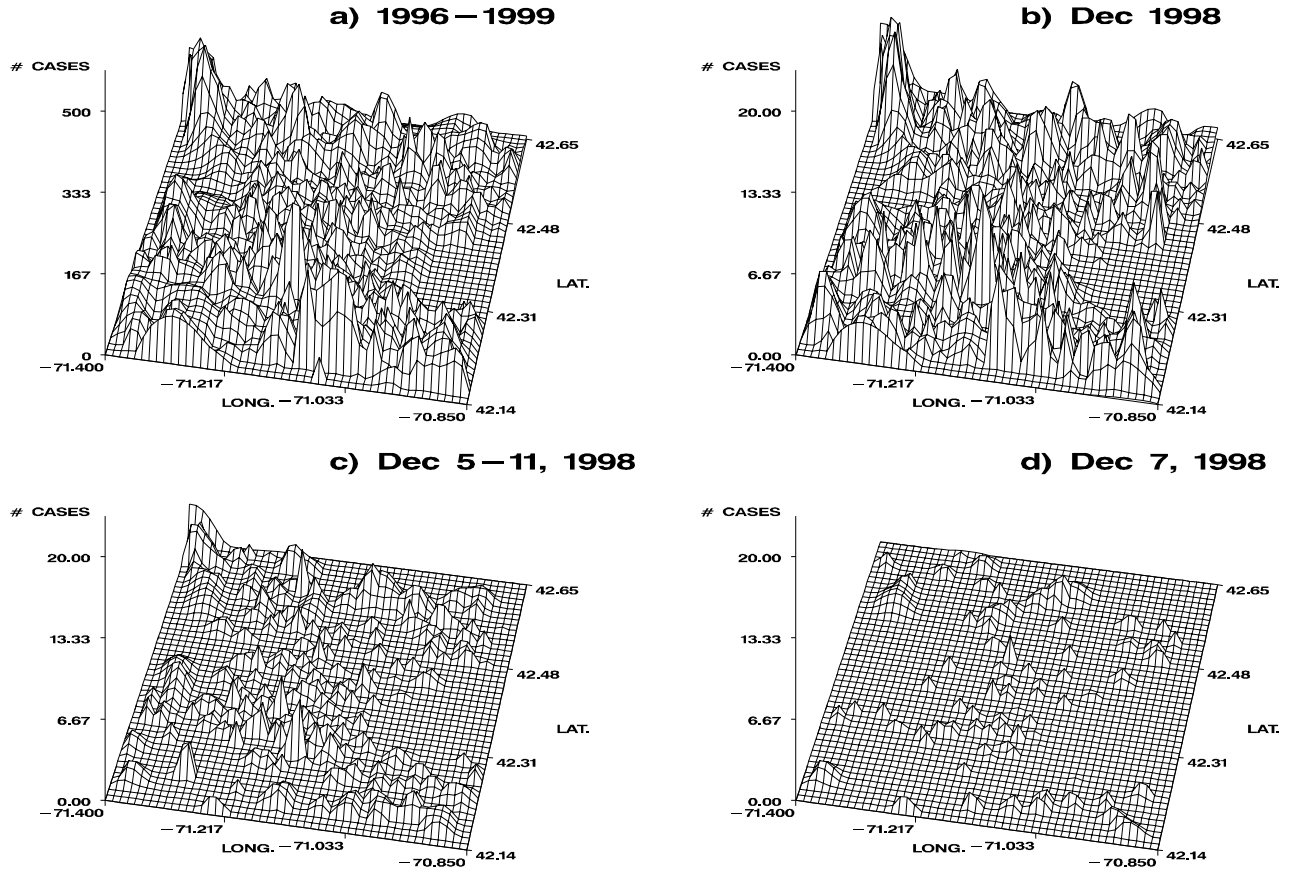


**Figure 5**
Distribution of lower respiratory infection episodes in time and space. Each episode is mapped to the census block group in which the individual resided. Four durations are shown: 4 years (upper left), one month (upper right), one week (lower left), and a single day (lower right). Note that scale of the vertical axis is different for the upper left panel, compared to the other three. The x-axis is longitude, and the y-axis is latitude. The area with no episodes on the right border of each plot is water. Counts are shown rather than rates to correspond to the data shown in Figures 3 and 4. Variations between census tracts are principally due to the distribution of health plan members' residence locations.

described by Armstrong [6] and de Wit [7], that have been created to collect information that isn't captured by standard reporting systems. Finally, data from automated medical record systems lend themselves well to incorporation in automated detection systems with little or no added cost, because the data are available in electronic form, avoiding the additional costs and errors due to data entry.

For a data source to be useful for disease surveillance, it must be timely, accurate, complete, and capable of distinguishing events of interest from background occurrences, i.e., an acceptable signal to noise ratio. The interval between a patient being seen and data being available for analysis must be short, particularly in any bioterrorist threat situation. Diagnostic and demographic accuracy are needed in order to enable reliable evaluation of geographic clustering of specific emerging infections or syndromes. Complete data or at least reasonably complete sampling is essential if events of relatively small scale are to be detected.

In terms of timeliness, this ambulatory record information can be available very quickly. In practice, it is efficient extract each day's visits of interest during the succeeding night, thus making the information available within 24 hours of the patient encounter.

In terms of accuracy, the data are probably of acceptable reliability for patient demographics and encounter dates, since this information comes from the administrative database used for reimbursement. The validity or reliability of physician diagnosis in terms of ICD9 codes is neither known nor readily amenable to measurement. In other

settings, ICD9 codes have been shown to have substantial discrepancies, when they are compared to the information in the full text medical record [8]. We expect that the diagnoses of interest here also have substantial errors that reduce both sensitivity and specificity. It is likely that these errors are relatively stable over the time periods of interest for surveillance of acute disease syndromes, and so this problem may not interfere appreciably with day to day comparisons. However, there could be important differences attributable to coding practices between groups of clinicians, or in different medical record systems, for instance if the automated systems guide clinicians to choose certain codes over others.

The lack of uniformity in the use of ICD9 codes, for instance in assigning a diagnosis of pneumonia, may be ameliorated by grouping diagnoses into broader syndromic surveillance categories, as was done for this report. It is notable that the relatively non-specific diagnosis of "cough" accounted for more than half of LRI encounters studied here. At this time, we have no simple way to measure directly the accuracy of coding or of directly assessing the effect of syndromic groupings. No standard syndromic grouping is yet in wide use for surveillance purposes. The provisional grouping used in this report was developed by the U.S. Department of Defense for its own specific needs but appears to have worked reasonably well with the ambulatory care data described here. Because a small number of relatively non-specific codes (cough, pneumonia, bronchitis) account for the large majority of episodes of lower respiratory infection, it is likely that most syndrome groupings will yield very similar results. Although the inclusion of symptoms like cough clearly reduces specificity, we believe this is outweighed by the gain in sensitivity. This tradeoff is discussed in more detail below.

In terms of completeness, an automated system like the one we describe here is typically as complete for ambulatory encounters as the records that clinicians maintain. Our belated discovery that some encounters of interest were missing from the analysis dataset adds an important cautionary note, however, about the potential problems in adapting data designed for one purpose to another one. We believe the omission of emergency room visits has a minor impact on the total number of events, since their number is small in relation to the total number of ambulatory visits. However, surveillance based in emergency rooms is also of great value. We see the system described here as being complementary to emergency room based systems. A potential advantage of assessment of office visits is the possibility that it will provide an earlier signal than will an emergency room based detection system if the condition of interest begins with symptoms that don't warrant emergency room care.

The use of data from an automated medical record system in a health care environment linked to individuals' insurance coverage provides an additional reason to believe the data are reasonably complete, since individuals have a strong financial incentive to receive their care from providers whose clinical encounter information is reported to the insurer for reimbursement purposes. It may also be possible to make similar use of diagnoses contained in automated billing data, rather than automated medical records. Although most current administrative systems include time lags that diminish their utility, the development of on-line transaction processing between clinicians and payers may reduce or eliminate that deficiency. All medical records or claims based systems depend, of course, on individuals' bringing the event to clinicians' attention. Such systems provide no information about the large number of illnesses that resolve without formal contact with the medical system.

Even large health plans typically include only a portion of individuals in a community. Although the number of individuals may be sufficient for surveillance purposes, it will be important to assess the degree to which the covered population resembles the entire population. Insured populations are likely to be adequate for many conditions of interest, especially if one adjusts for major determinants of illness.

An additional advantage of using automated data from health plans is the ability to know the exact size, composition, and residence location of the source population. Although we limited our characterization of the population to age and sex, it is also possible to use more detailed information about disease history, for instance to characterize the burden of illness among individuals with specific chronic diseases. Locating clusters of illness should have great utility for identifying and remedying localized disease sources; these might be locations, such as day care facilities, where infection is transmitted person to person, or areas in which there are environmental sources of infection, such as a contaminated water supply. The geographical information available to health plan is primarily useful, of course, for conditions whose source is near individuals' homes.

Primary medical records will include multiple encounters within a single episode of infection for some patients. The decision to define a new episode on the basis of six weeks free from any LRI coded encounter was based on a combination of clinical experience and the pattern of repeated encounters. For surveillance systems to be comparable, the classification of encounters into episodes of illness is an important methodological problem that deserves additional attention. It is possible, for instance, that the pattern of repeat visits might change

during a cluster of illness. Although the distribution of LRI visit intervals supports a six week disease free interval to become eligible for a second episode, other cutoffs might also have been chosen.

It seems reasonable to assume that if similar patterns of illness in time and space are observed in other systems (such as specific disease surveillance systems, hospital discharge records, or other large ambulatory care record systems), then this provides some degree of criterion validity for the data presented here. The striking similarity in seasonal variation between our LRI episodes and the national experience with pneumonia and influenza deaths provides one measure of assurance that our system identified relevant events. We also compared the CDC's pneumonia and influenza death data for Boston to our experience, but the number of reported deaths was too small for meaningful seasonal patterns to emerge. Additional support for the utility of medical record surveillance information comes from comparison of our data to that collected by the National Ambulatory Medical Care Survey (NAMCS) [9], which uses multistage sampling of ambulatory care physicians, requiring participants to report a random sample of patients seen in a randomly assigned week. For the period 1990 to 1996, the estimated rate for lower respiratory infection office visits was 74.2/1000 population per year, based on about 40,000 sampled records per year nationally. This estimate is reasonably close to our observed rate of 93/1000. The difference between the two rates may be due in part to sampling variation (principally arising from the smaller NAMCS sample), differences in the age/comorbidity profiles of the populations, greater sensitivity of the HVMA sample because it included telephone encounters, and lack of specificity of the cough diagnosis. In any event, the difference, even if real, should not seriously interfere with the utility of this syndromic surveillance system to identify overall disease trends or provide early warning of illness clusters.

Similarly, the age and sex distribution of these cases is consistent with our knowledge of the epidemiology of lower respiratory tract infection. Others, using data from an office practice, have shown an early childhood peak at approximately one year, also with higher rates in males of approximately 25% [10]. The increasing rate with age among adults has been widely recognized, along with an overall male predominance [11–13]. Some studies have reported either a smaller difference between men and women among younger adults [12], or an excess among younger adult women [11], as observed in our population. Our data do not distinguish between actual differences in disease incidence by age and sex, and differential ascertainment, either because of differences in likelihood of seeking health care or differences in the way clinicians code encounters for men and women.

In order to distinguish signals of interest from background occurrences, we believe it will be necessary to develop statistical methods to identify unusual clusters that deserve further attention. The volume of data acquired is so large that it is impractical to perform manual daily inspection of data from a large geographically dispersed population. This is especially important since there were only twelve lower respiratory infection syndrome clusters each year of more than approximately five events occurring in a single day among health plan members residing in a single census tract (authors' unpublished data). The specific number of events required to be included in the twelve most extreme clusters depended on the number of health plan members in a census tract, as well as the month of the year, and the day of the week.

The fact that relatively few events occurred on any particular day in any census tract supports our inclusion of the "cough" diagnosis in the syndrome definition, in order to improve the sensitivity of our case finding. Although this is a non-specific diagnosis, and it accounted for a majority of all events, the total number of these was not so large that it compromised the utility of this particular surveillance system. An enhancement that may be useful in automated medical record systems, but not in claims based systems, is to require fever (measured value, not ICD9 code) to be part of the definition of a lower respiratory infection. This would presumably preserve sensitivity for conditions like anthrax, and also reduce the number of false positive clinical events.

To the extent this surveillance method proves useful, it will be worthwhile to extend it to other conditions that cluster in the areas of residence of affected individuals. Within infectious diseases, these might include diseases spread by airborne dissemination in residential areas, by contaminated foods or water distributed to residents of a neighborhood, by insect or other animal vectors, or by person-to-person transmission in households (secondary spread). Specific infection syndromes of interest, in addition to lower respiratory infection, include upper respiratory infection, gastrointestinal disease, neurologic disease, and rash. It may also be useful for other conditions that may be clustered in time or space, such as injuries.

We conclude that as automated ambulatory care record systems become more widely available, they can assume an important, currently unfilled, role in disease surveillance. Such systems are less prone to undercounting than traditional public health reporting systems, and they are less resource intensive than traditional sentinel

surveillance systems. These data can serve several different purposes, including informing clinicians of conditions that are prevalent in their communities, providing detailed and timely information to health plans that need to allocate scarce resources, and to public health programs to allow early recognition and response to changing disease patterns. Suitably de-identified electronic data could be provided to public health systems in a format consistent with the emerging National Electronic Disease Surveillance System (NEDSS) standards [14]. Inclusion of such reporting capability, under clinicians' and health systems' control, in commercial medical record systems is likely to be an inexpensive way to provide the required data in the most usable form. The timely use of automated diagnosis information, especially with cluster detection algorithms, may be a valuable resource for supplementing current infectious disease surveillance systems.

## List of abbreviations

HMO: Health maintenance organization

HPHC: Harvard Pilgrim Health Care

HVMA: Harvard Vanguard Medical Associates

ICD9CM: International classification of disease, 9$^{th}$ version, clinical modification.

LRI: lower respiratory tract infection

NAMCS: National Ambulatory Medical Care Survey

NEDSS: National Electronic Disease Surveillance System

## Competing interests

None declared

## Additional material

### Appendix 1

*List of ICD9 codes contributing to the lower respiratory infection syndrome group.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2458-1-9-S1.doc]

## Acknowledgments

Supported by: Grant Award U90/CCU116997 from the Centers for Disease Control and Prevention / Massachusetts Department of Public Health Public Cooperative Agreement for Health Preparedness and Response for Bioterrorism.

We appreciate the help of Courtney Adams, of Harvard Pilgrim Health Care, in organizing and conducting this work, and the support and advice of Patricia Kludt, Mass. Dept. of Public Health, and Denise Koo, Claire Broome, and Robert Pinner, of the Centers for Disease Control, and Julie Pavlin, Department of Defense Global Emerging Infections System.

## References

1. Baxter R, Rubin R, Steinberg C, Carroll C, Shapiro J, Yang A: **Assessing core capacity for infectious disease surveillance.** *Final Report. Prepared for: Office of the Assistant Secretary for Planning and Evaluation, DHHS,. The Lewin Group, Inc* 20001-47
2. : **The International Classiciation of Diseases, 9th Revision, Clinical Modification. National Center for Health Statitistics.** *Published by Commission on professional and hospital activities, Ann Arbor, MI* 1980
3. Barnett GO, Justice NS, Somand ME, *et al*: **COSTAR-a computer-based medical informatioin systrem for ambulatory care.** *Proc IEEE* 1979, **67**:1226-1237
4. : **U.S. Department of Defense GEISaRS.** *Annual Report, Fiscal Year 1999: Walter Reed Army Institute of Research* 19991-32
5. **Centers for Disease Control and Prevention.** [http://www.cdc.gov/ncidod/diseases/flu/fluvirus.htm]
6. Armstrong GL, Pinner RW: **Outpatient visits for infectious diseases in the United States, 1980 through 1996.** *Archives of Internal Medicine* 1999, **159**:2531-2536
7. de Wit MA, Koopmans MP, Kortbeek LM, van Leeuwen NJ, Bartelds AI, van Duynhoven YT: **Gastroenteritis in sentinel general practices, The Netherlands.** *Emerging Infectious Diseases* 2001, **7**:82-91
8. Goldstein LB: **Accuracy of ICD-9-CM Coding for the Identification of Patients With Acute Ischemic Stroke: Effect of Modifier Codes.** *Stroke* 1998, **29**:1602-1604
9. National Ambulatory Medical Care Survey: **Public use data tape documentation [program]. Hyattsville, Md.: National Center for Health Statistics;** *Centers for Disease Control and Prevention* 1980
10. Murphy TF, Henderson FW, Clyde WA Jr, Collier AM, Denny FW Pneumonia: **An eleven-year study in a pediatric practice.** *Am J Epidemiol* 1981, **113**:12-21
11. Jokinen C, Heikanen L, Juvonen H, Kallinen S, Karkola K, Korppi M, Kurki S, Ronnberg P-R, Seppa A, Soimakaillio S, Sten M, Tanska M, Tarkiainen A, Tukiainen , Pyorala K, Makela PH: **Incidence of community-acquired pneumonia in the population of four municipalities in eastern Finland.** *Am J Epidmiolo* 1993, **137**:9778-988
12. Almirall J, Bolibar I, Vidal J, Sauca G, Coll P, Niklasson B, Bartolome M, Balanzo X: **Epidemiology of community-acquired pneumonia in adults: a population-based study.** *Eur Respir J* 2000, **15**:757-763
13. Marston BJ, Plouffe JF, File TM Jr, Hackman BA, Salstrom SJ, Lipman HB: **Incidence of community-acquired pneumonia requireing hospitalization. Results of a population-based active surveillance study in Ohio.** *Arch Int Med* 1997, **157**:1709-1718
14. The National Electronic Disease Surveillance System Working Group: **National Electronic Disease Surveillance System (NEDSS): A standards based approach to connect public health and clinical medicine.** *J Public Health Management and Practice* 2001